text extraction

Main plan: extract text and tables from a text file or html file. For the tables, we want to have a xlsx file including all the tables written in different sheets, the rows and columns of tables be well detected. For the text, we want to create a xlsx file including the page information, the type of the extracted text and the found text in the report.

Accomplished work:

We write the code to analyze the structure of the html code, find the <table> and <p> in the input file. We detect the <tr> and <td> in <table> so that we can get the contents of each cell separately. We also detect the style of the <p> to determine the type of paragraph and grab the page information according to the comment if html code.

Finally, we get the file we wanted.